AD-A254 711

DISCUSSION PAPER
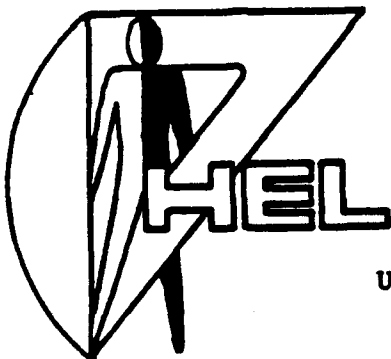
NO. 54

MODELING THE SPEECH COMMUNICATION EFFECT ON

PERFORMANCE: MESSAGE COMPLEXITY

June 1991

Andrew M. Rose
Donna J. Mayo
Janice C. Redish
American Institutes for Research

Deborah P. Birkmire
Leslie J. Peters
Georges R. Garinther
U.S. Army Human Engineering Laboratory

U.S. ARMY HUMAN ENGINEERING LABORATORY
Aberdeen Proving Ground, Maryland 21005-5001

92

MODELING THE SPEECH COMMUNICATION EFFECT ON PERFORMANCE:
MESSAGE COMPLEXITY

Andrew M. Rose, Donna J. Mayo, and Janice C. Redish

American Institutes for Research

Deborah P. Birkmire, Leslie J. Peters, and Georges R. Garinther

US Army Human Engineering Laboratory

ABSTRACT

This document describes a research plan which will lead to the development of a model that relates speech communication to performance. The experiments to be conducted answer questions about the effect of spoken message set complexity on operational performance. A message set has been defined as the set of all possible response-triggering messages within the boundaries imposed by the situation or setting. First-year and long-term objectives are presented. An experimental paradigm, independent and dependent variables, subject populations, and other details of the experiment are described. Operational definitions of the independent variables are presented. Additionally, a tentative research plan for continued studies is proposed.

DTIC QUALITY INSPECTED 3

Accession For

NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification

By

Distribution/

Availability Codes

Dist	Avail and/or Special
A-1	

TABLE OF CONTENTS

Modeling the Speech Communication Effect on Performance:

Message Complexity

INTRODUCTION

Speech communication is a critical component of effective military performance. In a very large number of individual tasks, and in an even higher percentage of collective (squad, crew, platoon, etc.) tasks, both the timing and the direction of the desired response is triggered by spoken communications. Thus:

* "First platoon; begin laying down covering fire when second platoon leaves its take-off position."

* "Battery A; fire for effect."

* "Following the countdown, begin a precision turn to heading 305; ...counting 5,4,3,2,1, NOW."

* "Driver back-up, gunner take over."

are examples of brief directive communications that are intended to produce a particular response at a specified time (in the second and fourth examples, "do it as soon as the message is received" is implied rather than made explicit). Rather obviously, if an effective response is to be triggered by a communication, the minimum requirement is that the message be received and comprehended.

Each of the above messages can be considered as a member of a message set, which we define as the set of all possible response-triggering messages within the boundaries imposed by the setting. Thus, the third message is an element of the set appropriate for the strike team of an aircraft; the fourth message is from the set appropriate to a tank crew in engagement. A message such as

"I have a three-day pass starting next Monday."

is not a member of a message set, since it is not a trigger for any operationally significant response. A message set can also be bound by a formal structure. For example, when a forward observer gives a report of an enemy sighting, the message content is structured by the SALUTE format (Size, Activity, Location, Unit or Uniform, Time, and Equipment). A message set also includes any codes, abbreviations, and idosyncrasies known to the people communicating.

A message set occupies a region of a general model of communication. The typical structural components of this general model are a transmitter or encoder, an encoding mechanism, a medium of transmission, a decoding mechanism, and a receiver. These structural components are activated by a message that must pass through each component. This message transmission always occurs in a particular

1

context, usually termed "background." Also, for any particular communication, an additional component is the action that is taken by the receiver following decoding: the behavior or response triggered by the message as decoded.

Messages undergo a series of transformations (encoding, decoding) as they proceed through the system; the efficiency of the system typically involves comparing the final decoding with the original message. However, another way to evaluate the efficiency of the system is to compare the response the message triggers to the response intended by the sender.

Each component of the system can add "noise" (in an informal, rather than engineering sense) to the transmission, either in the sense of degrading the final decoding or affecting the performance of the receiver. Using this terminology, we can state the basic problem to be addressed in this project: Can we quantify the "noise" added to the communication system by the message set, particularly in terms of its effect upon the resultant performance of the receiver?

The complexity of the message set can add noise at several points in the communication system. For example, increasing complexity can increase the difficulty - and therefore the probability of errors and increased transmission time - of encoding and decoding. It can also increase the demands upon the transmission medium.

Much of what we know about spoken communication effectiveness has been learned from military research. Most of the early research, and much of the current research, focuses on engineering aspects of communication, such as the development of methods for improving fidelity of the system and the development of more efficient coding strategies. This research considers the message itself principally as a series of acoustic signals; the message set, and the content of the messages, are examined primarily in terms of their physical aspects.

The other primary focus of research has been the reception of the message - that is, the decoding component. Much is known about the "comprehensibility" of a received message under varying mixes of signal and noise, and the effects of various aspects of message content (number of words, sentence structure, etc.) on message comprehensibility. However, little of this research has gone beyond measuring the simple comprehension of the message to relate this to operational effectiveness. A series of research studies (Peters & Garinther, 1990; Peters, Garinther, Birkmire, & Whittaker, in press) has related the effects of degraded speech on measures of mission success. In general, it was found that poorer comprehensibility resulted in longer operational times and a greater number of operational errors.

But researchers have also known that operational effectiveness - the responses resulting from decoding - could not be a simple function of either the engineering fidelity of the transmission or the comprehensibility of the message. Early attention was given to measures of "familiarity" of the elements making up a message; the development of restricted vocabularies was one response to the finding that comprehension scores were higher with restricted rather than unrestricted word sets. The concept of "predictability" is related; under most conditions, an expected message is received more accurately than an unexpected one.

2

Other variables have been studied; the influence of message set size is well-documented (e.g., Miller, Heise, & Lichten, 1951). It is an everyday observation that "known" speakers are better understood than strangers. One would assume that military teams that remain intact communicate more effectively than those that have experienced turnover, holding constant all other aspects of the communications. The context(s) within which a message is transmitted can affect both sender and receiver and hence the intelligibility of the message. The semantic confusion potential of a message set must affect the error rate in decoding. There are a large number of variables that can potentially effect message set complexity.

Increasing message set complexity may induce all sorts of performance problems having serious consequences in the operational environment. But again, there has been very little research concerning the relationship of any of these variables to operational effectiveness. Birkmire, Peters, and Stouffer (1990) measured conceptual density and redundancy for auditory communications in a gunnery exercise. They found very high correlations between their measures of communication complexity and target identifications. Additionally, Federman and Seigel (1965) showed that various measures of communication were predictive of mission accomplishment in ASW helicopter teams. But the independent variables used in this study were measures of the nature of the communication, that is whether it was invitational, evaluative, etc., rather than of semantic or syntactic content. Although it seems apparent that confusion potential, message length, and intelligibility are necessarily related to performance in the extreme and probably related over much of their scales, there is little supporting research. Further, the literature provides very little information regarding how such variables relate to one another, and, as a consequence, how they might be combined to provide an overall measure of message set complexity.

This document describes a research plan for experiments to be conducted by the American Institutes for Research (AIR) designed to answer some of these questions on the relationship of message set complexity and performance. We will present our first-year and long term objectives. Next, we describe the experimental paradigm, independent and dependent variable, subject, and other details of the experiment. Finally, we present a tentative research plan for continued studies.

Objectives

The overall, long-term objective of our program of research is to develop and test a model that relates message set complexity to performance. This program consists of three major phases, only the first of which is supported by the current contract. The major objectives of the first phase are as follows:

1. The first objective is to identify a set of "message" variables that we hypothesize would affect operational performance.

2. Once these variables have been identified, the second objective is to develop an experimental paradigm for further exploration of these variables.

3

3. The third objective is to evaluate the variables in light of two primary questions.

   o Can we define and measure each variable as a characteristic of operational messages? In other words, can the variable be applied to realistic (or all) messages?

   o Can we define and measure each variable in the context of messages to be used in the experimental paradigm?

4. The fourth and final objective of the first phase is to demonstrate the feasibility of the paradigm for investigating the relationship between the variables and performance. Can we collect meaningful performance measures that vary (at least generally) with changes in levels of the message characteristics? Can the experimental paradigm support studies of one-way (i.e., communication that commands or instructs), and two-way (i.e., communication that interrogates) communications where the level of speech intelligibility is varied? This is basically an "equipment check" to see if we can actually collect meaningful data that, at least heuristically, would support further experiments.

The major objective of the second phase of the research program is to actually develop a quantitative model of message complexity through a series of laboratory experiments. The basic approach will be to use the experimental paradigm developed in the first phase to collect performance data on a large number of messages. Each of the messages will have a score for each of the message variables identified in the first phase. Since the variables in the model are too numerous to support an Analysis of Variance design, that is, to parametrically vary all levels and combinations of variables, we will use a multiple regression approach. We will construct messages that contain all levels of the variables; the measures of these variables will serve as predictors of performance. If successful, this strategy will result in weights for each of the variables that will allow us to generate performance predictions for other messages. We will evaluate the goodness of fit of the model to the performance data, as well as the ability of the model to predict performance for new messages.

A secondary objective of the second phase of the research program is to extend the model to multiple-path communication tasks, tasks where it is necessary for two or more people to exchange information in order to produce some measurable performance.

If successful, the result of the first two phases of the research program will be a laboratory-validated model. The third phase of the research program will be to integrate our research with other ongoing HEL-sponsored work, and to validate the model against real-world, operational performance. We hope to conduct experiments and observations of people performing actual communication tasks with measurable performance requirements. We will use the model to

4

generate predictions, then evaluate the accuracy and reliability of those predictions. Whenever possible, the communication tasks will involve several controlled and measured levels of speech intelligibility. We envision experiments utilizing high-tech simulation systems (e.g., SIMNET) and systematic observations and analyses of actual performance in military training situations.

PHASE ONE EXPERIMENTAL PLAN

Objective One: Identify Message Variables

As a result of our review of relevant research and discussions between HEL and AIR, we have derived an initial set of message variables that appear, both empirically and theoretically, to affect comprehension of written text. As far as can be determined, however, empirical support is lacking for several critical aspects of these variables. First, there has been little or no experimental work that establishes the relationships among these variables, especially when applied to spoken communication. In addition, there is little or no experimental work involving the interaction of any of these variables with levels of speech intelligibility, particularly given the way in which we will implement the degradations.

The current list of message variables is the following:

- message length
- number of ideas
- word frequency
- redundancy
- morphological confusion
- given-new vs. new-given order
- expectancy
- passive vs. active
- stative vs. action verb
- personal vs. impersonal
- nominalization vs. action verb
- levels of subordination
- type of branching for subordination

We have reduced this list by combining some variables from an earlier list of 15. Since "average clause length" is a combination of "number of ideas" and "message length" it was dropped from the list. "Subordinate vs. simple" is a dichotomous variable that is now included as part of the "Levels of subordination" variable.

We will define variables below in our discussion of Objective Three. However, to repeat, during this first phase of experiments we will not determine precise quantitative relationships among these variables or between each of these variables and performance; rather, we will use the experimental results to determine the feasibility of using these variables in the context of our experimental paradigm. This involves operationally defining each variable, both for general messages and for the specific context of our experimental paradigm.

Objective Two: Develop an Experimental Paradigm

The second objective is to develop a paradigm to experimentally study message complexity. We will conduct these experiments while independently

varying the level of speech intelligibility. We will conduct these experiments within two communication "structures," namely one-way, or command, and two-way, or interrogation, configurations. A third configuration, multiple-path communications (e.g., discussions) will not be addressed in this first round of experiments. However, if the current paradigm proves to be usable with the other structures, we would be able to adapt it for the more complicated situation.

To illustrate the basic experiment, we first describe a typical task from the subjects' perspective. We then discuss the specific details and variations planned for this phase of the research.

One-Way Communication. In this paradigm, two subjects--the speaker and the respondent--are seated in two different rooms and communicate through an intercom system. The speech intelligibility from the speaker to the respondent is controlled by passing the speech signal through a circuit which chops speech at a rate of 60 Hz and varies the on-off duty cycle of the speech signal between 100% (i.e., perfect intelligibility) and 0% (no intelligibility). The settings for the chopping circuit are under the control of the experimenter and are calibrated for the two subjects. In the course of an experimental session, the respondent will hear messages at four levels of intelligibility.

Both subjects face identical computer displays: an eight-by-eight grid, wherein each square is one of four colors (for example, blue, green, red, and yellow). A cursor appears in one of the squares; only the respondent controls the movement of the cursor by pressing the arrow keys on a keyboard. As an option in the program, the speaker's display can show where the respondent moves the cursor; however, we do not plan to use this option during the first phase of the research program.

A typical trial would proceed as follows:

1. The speaker presses the message button which also turns on the microphone.

2. A message appears on the speaker's screen. A typical message might be, "Move three squares north to the second yellow square."

3. The speaker reads the message into the microphone.

4. The respondent moves the cursor according to the directions heard in the message.

5. When the move is complete, the respondent presses a button to indicate readiness for the next message.

6. The speaker then presses the message button, and the next message is displayed.

A problem ends when the respondent finds the square that reveals the message, "End," or when the experimenter terminates it via the computer.

Two-Way Communication. The basic difference in the typical trial for two-way communication is that the speaker must interpret the respondent's message

about the position of the cursor. Based on the interpretation of the position report, the speaker moves the cursor to the understood position. This position determines the speaker's next message. Thus, in a two-way communication trial, both the speaker's and the respondent's messages are scripted.

A typical trial would be as follows:

1. The speaker presses the message but.on.

2. A message appears on the speaker's screen. A typical message might be, "Move three squares north to the second yellow square."

3. The speaker reads the message into the microphone.

4. The respondent moves the cursor according to the directions heard in the message.

5. When the move is complete, the respondent presses the message button.

6. A message appears on the respondent's screen (e.g., "I am three squares south of the blue square").

7. The respondent reads the message into the microphone.

8. The speaker moves the cursor on the display according to how the respondent's message is interpreted.

9. The speaker then presses the microphone button, and the next message is displayed.

Again, the problem ends when the respondent finds the square that reveals the message, "End," or when the experimenter terminates it via the computer.

Apparatus. Two SSI Microfocus 386 System computers with ViewSonic 4 color monitors have been configured for the experiments. The two computers are linked through serial communications ports. Two modified response pedestals are to be used as movement interfaces during the experiment. They are capable of being tied into the computers so that a designated button on each pedestal displays the message and keys the microphone while activating the timing circuit in the computer. Also, the capability exists of using either a joystick on the response pedestal or the computer keyboard to move the cursor during the experiments. Additionally, two microphones, an amplifier, and headphones which are connected through the chopping circuit have been configured for the experiment.

The display and problem presentation. Each "problem" is designed around one of several predetermined paths through the grid, leading from a start square to an end square. Messages are scripted for each square that direct the respondent along (or back to) the desired path. We have designed a flexible display that can present a grid with any number of rows and columns and with any of sixteen colors in a square. We have also drawn a series of icons (e.g., a bridge, a house, a barn, etc.) that can be placed anywhere on the display.

7

Once colors and messages are assigned to squares, we can rotate, invert, and flip-flop the display to produce a large number of "equivalent" problems. These transpositions maintain the characteristics of the predetermined paths, thereby controlling the difficulty of the required movements. In other words, transposing a problem maintains all movement distances and number of turns, while producing an entirely "new" problem for the subject.

We have created six basic paths that we will use to develop the individual problems for the initial experiments. Improvements to the computer program have increased the flexibility that we have in designing problems. We have created specification files for each of the basic paths using a word processing program. We are in the process of transferring these files into a "design" program which automatically assigns colors, messages, and path directions to each square in a matrix. Once this activity is completed, generating the problems for the experiment will be accomplished quickly. We have also developed a "testing" program that we can use to verify that problems are functioning correctly.

In order to run the experiments, we created an "executable" program. This program allows us to write a file for each experimental session that will automatically present a predetermined series of experimental problems. This program is very flexible; it allows us to present problems in any order, to present instructions to the subjects on the screen, and to pause at designated times during the sessions. This program also writes the responses of the subjects to data files.

Responses and dependent variables. Each problem consists of a series of trials. A trial is defined as the sequence of events listed below. The "executable" program will time these events for each trial:

1.   speaker reads/transmits message

2.   respondent receives/interprets message

3.   respondent moves cursor

3a.   -- response latency between respondent's
        final final cursor movement and
        transmitting the response

4.   respondent read/transmits response

5.   speaker receives/interprets response

6.   speaker moves cursor [two-way communication
       only]

6a.   -- response latency between speaker's final
        cursor movement and transmitting the next
        message [two-way communication only]

Although the response latencies are not relevant to the experiment, we included these events in order to create a complete time record for each trial.

8

The program has also been designed to record data on the accuracy of the speaker's and respondent's cursor movements. Within each trial the computer will record:

- correct/incorrect cursor movements--whether the correct target square was achieved for each message

- the absolute error of the cursor position--the number of squares from the final cursor position to the correct target position

More globally, we will record the total time and the number of messages required to complete the problem.

Objective Three: Define and Operationalize the Message Variables

Given the list of message variables and the experimental paradigm described above, the next objective is to make the variables more concrete. We define the variables below, and give examples of how we operationalize each as speaker messages in the experiment.

Message length

> **Definition:** the number of words in the message
>
> **Levels:** continuous; a simple count of the total number of words; alternatively, short (7 or less words), medium (8-16 words), and long (greater than 16 words)
>
> **Examples:** Go east to the first brown square. Go south to the next brown square. (14 words)
>
> Go east to the first brown square and then go south to the next brown square. (16 words)
>
> Head east and turn south at the first brown square, continuing until you come to another brown square. (18 words)

Number of ideas

> **Definition:** the number of moves and supporting ideas in the overall message
>
> **Levels:** continuous; one, two, or three moves, plus supporting ideas
>
> **Examples:** 1 move: Go north to the first red square.
>
> 1 move,
> 1 supporting: Go north to the first red square by the barn.

9

2 moves: Go north to the first red square,
                      then go east to the second green
                      square.

    3 moves: Go north to the first red square,
                      then go east to the second green
                      square, then go south to the first
                      blue square.

Word frequency

   Definition: the commonness of a word in context

   Levels: common and not common (i.e., the message contains or
               does not contain any uncommon words)

   Examples: a. Go west to the second red square.
                    b. Advance west to the second scarlet square.

                    a. Go to the barn.
                    b. Proceed to the animal shelter.

                    colors: yellow-lemon, blue-aquamarine,
                            red-scarlet, green-emerald

                    icons: house-residence, bridge-overpass.

Redundancy

   Definition: extra information that repeats or restates an
                    idea

   Levels: none, one redundant element, two redundant
               elements

   Examples: a. Go north three squares.
                  b. Go north three squares to the red square.
                  c. Go north three squares to the red square near the
                     barn.

Morphological confusion

   Definition: a consonant cluster that is the same or similar in
                    sets of functionally equivalent words

   Levels: no confusion, potential confusion

   Examples: colors: blue, black, brown, bronze squares

                  icons: barn, bridge, bike, building trees, troops,
                          trucks, tanks

10

Given-new vs. new-given order

    Definition: The goal is presented before or after

    Levels: the instruction goal first, goal last

    Examples: You are going to the house. Go north to the second blue square. Then go east to the yellow square.

                Go north to the second blue square. Then go east to the yellow square. You are going to the house.

                Starting at the hills, go the house along the path with the fewest brown squares.

                Taking the path with the fewest brown squares, go from the hill to the house.

                When you arrive at the next green square, go south.

                Go south when you arrive at the next green square.

Expectancy

    Definition: the degree of structural consistency across a set of functionally similar messages

    Levels: consistent, inconsistent

    Examples: Consistent:

                a1. Go north three squares to the first red square.

                a2. Go south two squares to the second blue square.

                a3. Go east four squares to the first purple square.

                Inconsistent:

                b1. Go to the first red square, three squares north.

                b2. Go two squares south to the second blue square.

                b3 Go east to the first purple square four squares way.

Passive vs. active

    Definition: the subject of the sentence is the object of the actor. The verb is correspondingly passive or active.

    Levels: passive or active

Examples: Passive:

  The hills are approached from the barns two red
  and then two green squares are passed.

  Active:

  Approach the hills from the barn. Pass two red
  squares and then two green squares

Stative sentence vs. action sentence

Definition: whether the sentence states a fact (verb = "is"
  or "are") or suggests an action (verb =
  imperative)

Levels: stative with one action sentence or all sentences are
  action sentences (imperative)

Examples: Stative:

  There is a brown square by the barn. There is a red
  square north of the brown square. There are two green
  squares between that red square and the hills. Take that
  path to the hills.

  Active:

  Start on the brown square by the barn. Go to the red
  square north of the brown square. Stop on two green
  squares between that red square and the hills. Take that
  path to the hills.

Personal vs. impersonal

Definition: whether or not the pronoun "you" or "your" is
  stated or implied in the sentence

Levels: personal or impersonal

Examples: Personal:

  You are on the red square west of the truck. You
  must get to the house. Stop at three brown squares
  on your way.

  Impersonal:

  There is a red square west of the truck. The final
  stop is the house. Three brown squares are the
  stops along the way.

Nominalization vs. action verbs

Definition: whether or not verbs have been turned into nouns

Levels: nominalization present or action verb present

Examples: **Nominal:**

Advancement to the tree is required.Make a southerly adjustment upon arrival at the next green square.

You are trying to reach the tree. Passage through one black square and then two red squares is necessary for achievement of the destination.

Action:

Advance to the tree. Turn south when you arrive at the next green square.

You are trying to reach the tree. You have to pass one black square and then two red squares to get there.

Number of levels of subordination

Definition: how many clauses are subordinated and how many are nested in a subordinate clause

Levels: score both number of subordinations and number of nestings.

Examples: None (simple sentences):

Start at the easternmost red square. Go west. Turn south at blue. Then turn north at green. Stop at the next brown square.

One subordinate clause - no further dependent clauses within it:

Starting at the easternmost red square,turn west.

Two levels of subordination - one dependent clause within a subordinate clause:

Starting at the northernmost green square that is next to a brown square, move two squares left.

Starting at the easternmost red square, turning at red, then brown, then green squares, go to the next brown square.

13

Three levels of subordination - a dependent clause
on a dependent clause in subordination:

Starting at a black square that is next to a red
square that is at the southern edge of the area,
move two squares north.

Type of branching for subordination

Definition: Where the subordinate clause appears: before,
after, or in the middle of the main clause

Levels: none, left, right, or center

Examples: Right:

Go to the house from the barn following the path
with three black squares.

Left:

Following the path with three black squares, go to
the house from the barn.

Center:

Go the house following the path with three black
squares from the barn.

Conceptually, these variables are independent: it should be possible to
construct messages with all combinations of all of the levels of each variable.
There is at least one obvious exception: if a message has no subordinate clauses,
it cannot be left or right branching. We have not attempted to examine all
possible combinations (with twelve variables, with only two levels of each, there
are 2^{12} = 4096 combinations). It is also true that many of these variables
covary naturally in real messages--for example, more redundancy in a message will
usually mean a longer message. On the other hand, we are confident that we can
assign values of all these variables to any message. these considerations will
be important for the second and third phases of the research program.

Objective Four: Collect Preliminary Feasibility Data

The fourth objective is to "put it all together": to collect data from
experimental trials that will provide heuristic information about the variables,
the experimental paradigm, and directions for work during the second and third
phases of the research program. Below, we describe some basic experimental
considerations.

Independent variables. The main independent variables are the content of
the messages and the level of speech intelligibility. During the first phase of
the research program, we plan to use all of the message variables in constructing
the scripts. Our current plan is to use four levels of intelligibility: 25%,
50%, 75%, and 100%, as has been used by other researchers in the HEL program
(Peters & Garinther, 1990).

Responses and dependent variables. With each trial, the computer will record all time intervals:

-- time to read and transmit the speaker's message

-- time between the end of the speaker's transmission and the beginning of the respondent's movement

-- respondent's movement time

-- time to read and transmit the responder's message

-- time between the end of the respondent's message and the presentation of the speaker's next message

The computer will also record the respondent's actual cursor movements. These movements will be classified as correct or incorrect; in addition, we will record the absolute "error" of the movement. More globally, we will also record the number of messages needed to reach the "end" square, and the total time to complete the problem.

Subjects. We will use six AIR staff members, none of whom work on this project, as subjects. These six people will be combined to produce 30 pairs of subjects. Each of these pairs will be "calibrated" on the chopping circuit using the Modified Rhyme Test for four levels of intelligibility.

In our judgment, the advantages of using the same subjects in the different experiments and in different roles within the same experiment outweigh the possible disadvantages. All subjects will serve as speakers and respondents in combination with all other subjects in each condition. Since the primary purpose of the experiment is to demonstrate the feasibility of the experimental paradigm, we will not be conducting the types of statistical analyses where the multiple use of each subject would affect the usefulness of the results. This use of subjects will minimize the effects of learning and practice, thereby reducing the variance associated with these factors. This also adds to the efficiency of the data collection: we will not have to explain the paradigm to new subjects each time we change the conditions or the experiment. Finally, we will be more confident of our subjects' motivation and effort. Since we will not repeat a specific problem for any pair of subjects, there will be no possible effect of subjects' recalling a path or specific set of moves.

Experimental design considerations. Again, since the primary purpose for conducting the experiments is as a feasibility demonstration, we do not plan an extensive analysis of the performance data. We will, of course, carefully control the experimental situation to avoid any unwanted effects. Our plan is to conduct each experiment as a completely within-subject design. All "subjects" (i.e., speaker-respondent pairs) will receive all levels of each message variable at each level of intelligibility.

We plan to run approximately 48 problems per "conditions"--that is, where one particular message variable is varied systematically. For example, suppose the "focal" variable is redundancy, which as described above has three levels. Each subject pair would receive 48 trials: four at each level of redundancy at

each of the four levels of intelligibility. In addition to the redundancy levels, each message will be assigned a score for each of the other message variables. Basically, we will present as many trials as possible within the constraints of time, expense, and subject fatigue.

Analyses. When conducting a feasibility study, the main question that is asked is whether the materials--the variables, the computer program, the experimental procedures--do what they are supposed to do. In the present context, we will address several issues during the test trials.

- Can we use the operational definitions of the message variables to systematically generate experimental stimuli? The examples described above demonstrate that, considered individually, we can generate message with specific attributes. The critical issue, however, is whether we can generate message for _any_ combination of attributes. Similarly, can we take any specific message and assign values to all of the message variables?

- Does the equipment work? What are the problems associated with synchronizing two computers, timing circuits, microphones, headsets, and chopping circuits?

- Does the computer program work? Are stimuli presented with no discernible delays? Does the computer operate rapidly enough to allow for accurate recording of time and errors?

- Are the tasks we envision for speakers and respondents reasonable? Is even the simplest condition impossible to perform at 25% intelligibility? Will the mechanics of performance (e.g., pressing buttons to obtain messages) interfere with the data collection? How long can subjects perform the tasks before they become fatigued?

- To the extent that it is possible to determine, does the program obtain accurate and reliable data? While some things are easy to check (e.g., the number of moves in a problem, the number of incorrect moves, the distance moved, etc.), the accuracy and reliability of the timing mechanisms is more difficult to determine

- As a subjective judgment, do the experimental results reflect the anticipated effects of the message variables? For example, do longer messages result in more errors or increases in response time? Are morphologically confusing words confused more frequently when intelligibility is reduced?

PHASE TWO PLANS

Overview

The major objective of the second phase of the research program is to develop a quantitative model of message complexity through a series of laboratory

16

studies. The basic strategy will be to use a multiple regression approach. We will use the experimental paradigm developed in the first phase to collect performance data on a large number of messages. Each of the messages will have a score for each of the message variables identified in the first phase. We will construct messages that contain combinations of the different levels of the variables; the measures of these variables will serve as predictors of performance. If successful, this strategy will result in weights for each of the variables that will allow us to generate performance predictions for other messages. We will evaluate the goodness of fit of the model to the performance data, as well as the ability of the model to predict performance for new messages.

A secondary objective of this phase of the research program is to extend the experimental paradigm to multiple-path communication tasks: tasks where it is necessary for three or more people to exchange information in order to produce some measurable performance. This extension of the experimental paradigm has implications for model construction: it may be that we will have to construct different models for the different communication contexts (i.e., one-way, two-way, and multiple-path).

An additional objective is to examine the model in light of the effects of changes in speech intelligibility. The issue is whether and how the model predicts changes in performance at different levels of intelligibility; again, it may be that variable weights must be changed or a completely different model would be necessary to provide better predictions.

Before discussing our specific plans and activities, we would like to clarify an issue that has arisen during the current phase of the research. During the course of the project, we have stated two somewhat incompatible research objectives. One is that of testing specific hypotheses regarding levels of selected independent variable--that is, testing whether or not a particular independent variable is "useful" in accounting for variance in the dependent variable. The second state objective is that of constructing a predictive model that predicts performance based on measurable attributes of a message and the communication situation.

Consider first the notion of hypothesis testing. Assume that there will be fourteen independent variables--the twelve message variables, plus level of intelligibility, plus communication structure. In order to exhaust all hypotheses regarding each variable and its interactions with all other variables, it would be necessary to collect an infeasible amount of data. To construct a series of experiments with at least two levels of each independent variable would require 16,384 messages to realize all combinations. Moreover, in order to obtain at least one degree of freedom in each cell for hypothesis testing, at least twice that number of messages would have to be presented to subjects. This is not particularly disconcerting, because a very large number of the potential hypotheses are either uninteresting or of no practical importance. For example, any significant interactions higher that first or second order would probably be uninterpretable. If hypothesis was the only focus of the research, the practical issue from this hypothesis-testing perspective would be to "surgically" postulate relationships that could contribute to the goal of predicting performance.

The primary issue from the perspective of model building is in the sampling of levels of the independent variables. The practical problem is that of

17

obtaining a predicted value for an observation drawn randomly from a defined population. The population, of course should be that from which observations would be drawn for practical application. It is arguable as to what the appropriate population should be. From a strictly theoretical model-building objective, one possibility is to consider the 16,384 combinations of independent variable levels as a population and randomly sample from it. Another strategy is to assess what conditions may arise in practical (e.g., military) operations and construct a population based on the target population of messages, and develop a sampling plan accordingly.

It is not our intent to conduct the second phase of the research program in the "hypothesis testing" mode. To repeat, the major objective of the research program is to develop a predictive model of performance, based upon the attributes of messages. We believe that systematic hypothesis testing is not compatible with developing a model within a realistic time frame.

Even from a model-building perspective, however, it is exceedingly important to evaluate which independent variables are most useful before collecting data. The number of "predictors" bears directly upon the expected validity of the estimate parameters of the prediction model. That evaluation has been and will continue to be an important consideration in our research. In a nontechnical sense, we think it exceedingly important to "understand" the model from a theoretical perspective in order to maximize its usefulness and applicability.

Basic Plan

Independent variables. In one sense, this phase of the research program is straightforward: the objective is to construct a predictive model that permits the estimation of dependent measures from a potentially large set of independent variable conditions, a fairly common research problem. The typical approach is to collect a large number of observations of the dependent measures and to conduct multiple regression analyses. This approach has been successful in developing selection tests, predicting skill decay rates for military tasks, and in developing predictive models of training effectiveness, to name only a few examples that AIR has accomplished.

In the simplest case, "level of intelligibility" and "type of communication task" would be additional predictor variables. The plan would be the same as before: to collect a large number of observations at different levels of these variables, and to enter them in a regression analysis.

However, the actual situation may be complicated if the effects of these two variables affect (or interact with) the basic nature of the effects of the other predictor variables. For example, if reduced intelligibility fundamentally changes the cognitive requirements of the listener--rather than simply making the processing more difficult--the established weights of the predictor variables may change.

To illustrate, several years ago we examined the prediction of performance in an electronic fault-finding task, using individual abilities as predictors (Rose, Fingerman, Wheaton, Eisner, & Kramer, 1974). We determined which factors accounted for performance as a function of changes in task characteristics. In the basic form of the task, we found that the factors of Flexibility of Closure/Spatial Scanning, Syllogistic Reasoning, and Memory all predicted

18

performance. However, as we changed task characteristics by making it more difficult, that is, by increasing the number of potential fault in an electronic circuit, _different_ abilities (Perceptual Speed, Inductive Reasoning) predicted performance. We concluded that "...different abilities are involved, and at different levels of involvement, when either the task dimensions vary or different measures are examined." Thus, the fundamental nature of the task changed; in the current situation, the implication is that different models may be necessary if intelligibility or communication situation has the same sort of effects.

Constructing test messages. The basic plan is to construct and test a large number of messages in various conditions of speech intelligibility and communication structure. The critical questions to be addressed in the second phase are what messages and how many messages to construct. As mentioned above, there are two basic strategies for addressing these question

> We can randomly select values for each message attribute and construct messages in accordance with the random selection. For example, we would randomly select "Short message, two ideas, no uncommon words, one redundant element,..." as the variable values and construct an appropriate message.

> We can systematically select message attributes. This could take the form of a stratified sampling plan (e.g., select two attributes to vary parametrically and sample other attributes randomly with those conditions), or a statistically-driven sampling approach such as a partially balanced incomplete block design. Another way to systematically select message attributes is to derive them from some population of interest, such as actual military messages. For example, we could take existing message sets from current Army doctrine, determine the common attributes of these messages, and construct messages according to these parameters.

Selecting between these strategies is both a statistical and a practical decision. We estimate that at current levels of staffing and budget, we could develop approximately 200 messages, and test them (repeated measures) for 30-50 subjects. Given this constraint, we believe that a systematic sampling plan would be a more realistic strategy than random sampling, especially if we focus on "real" messages.

Analysis strategies. Building a predictive model invokes a number of statistical issues not ordinarily considered in laboratory experimentation. One such issue is the "colinearity" introduced by imbalances of effect representation, both because of sampling and because the effects may not be orthogonal in the population from which the sample is drawn. Because of the nonorthogonality, the "effects" share common variance in the dependent variable, making it difficult and often unreasonable to make statements about the magnitude of effects based upon the estimated parameters of the prediction function. This is sometimes referred to as the "bouncing beta" phenomenon.

Another issue is that of "shrinkage" of validity when the prediction function is applied to new observations drawn from the population of interest. Shrinkage occurs because the variables of the model are, by sampling definition, _random effects_. Thus, application of least-squares optimization in a

19

multivariate model leads to overestimation of the correlation between predicted scores and the dependent variable. Another aspect of the shrinkage phenomenon is that the expected validity (predicted vs. actual score correlation in the population) changes with the size of the sample drawn from the population. Thus, the parameter estimated in validity studies is, in part, actually a function of the sample size.

Shrinkage is reduced in two ways: increasing the sample size, and reducing the number of variables. Increasing the sample size, that is, the number of observations per message, improves both the estimates and the parameter. Decreasing the number of variables reduces the opportunities for "unfortuitous" imbalances in sampling that lead to error in estimation. Statistical assessment of shrinkage is usually accomplished in one of two ways: (1) application of "shrinkage" formulas, such as those proposed by Lord (1950), Nicholson (1960), etc., or (2) by sample-splitting, cross-validation designs patterned after the Mosier (1951) double cross-validation logic. Many useful variants of sample-splitting strategies are possible, where the computer makes many random splits (see, e.g., Rosse, 1974).

Conclusions

Given all of the above considerations, our current plan reduces to the following activities:

1. We will determine a message sampling strategy. We currently favor selecting message sets now in use, plus selecting additional messages that "improve" the sample of message variables by avoiding colinearities.

2. We will develop a set of approximately 50 messages, based upon a subset of the message variables in accordance with the sampling plan.

3. We will administer the messages to approximately 30 pairs of speakers-listeners, collecting appropriate performance measures.

4. We will analyze the results, using multiple regression, to determine an initial set of beta weights and combination rules.

5. We will generate additional sets of messages that will serve to first test the "current" model, and second to recalculate beta weights.

6. We will repeat the above steps until we have tested approximately 200 messages.

We expect that this process will produce a high quality model relating message set variables to performance on operationally significant tasks.

REFERENCES

Birkmire, D.P., Peters, L.J., & Stouffer, J.L. (1990, November). _Predicting task performance in unfavorable signal-to-noise environments_. Presented at the annual meeting of the American Speech-Language-Hearing Association, Seattle, Washington.

Federman, P. & Siegel, A. (1965). _Communications as a Measureable Index of Team Behavior_. Technical Report: NAVTRADEVCENT 1537-1. Orlando, FL. US Naval Training Device Center.

Lord, F.M. (1950). _Efficiency of Prediction When a Regression Equation from One Sample Is Used in a New Sample_. (ETS RB 50-40). Princeton, NJ: Educational Testing Service.

Miller, G.A., Heise, G.A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. _Journal of Experimental Psychology_, _41_, 329-335.

Mosier, C.I. (1951). Problems of designs of cross validation. _Educational and Psychological Measurement_, _11_, 5-11.

Nicholson, G.E. (1960). Prediction in future samples. In I. Olkin, et al. (Eds.), _Contributions to Probability and Statistics_. Standord, CA: Stanford University Press.

Peters, L.J. & Garinther, G.R. (1990). _The Effects of Speech Intelligibility on Crew Performance in an M1A1 Tank Simulator_. (TM 11-90). Aberdeen Proving Ground, Maryland: US Army Human Engineering Laboratory.

Peters, L.J., Garinther, G.R., Birkmire, D.P., & Whittaker, L. (In press). Tactical implications of the effect of degraded communication on tracked vehicle operations. _Army RD & A Bulletin_.

Rose, A.M., Fingerman, P.W., Wheaton, G.R., Eisner, E., & Kramer, G. (1974). _Methods for predicting job-ability requirements: II. Ability requirements as a function of changes in the characteristics of an electronic fault-finding task_. Washington, DC: American Institutes for Research.

Rosse, R.L. (1974). _Estimation of Cross-Validity through Monte Carlo Sample Splitting: A General Approach_. Minneapolis, MN: Personnel Decisions Research Institute.